TION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 1993	Professional paper

4. TITLE AND SUBTITLE

A HIERARCHICAL CLUSTERING NETWORK BASED ON A MODEL OF OLFACTORY PROCESSING

5. FUNDING NUMBERS

PR: EE20
PE: 0601153N
WU: DN301044

6. AUTHOR(S)

P. A. Shoemaker, C. G. Hutchens, S. B. Patil

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Command, Control and Ocean Surveillance Center (NCCOSC)
RDT&E Division
San Diego, CA 92152-5001

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

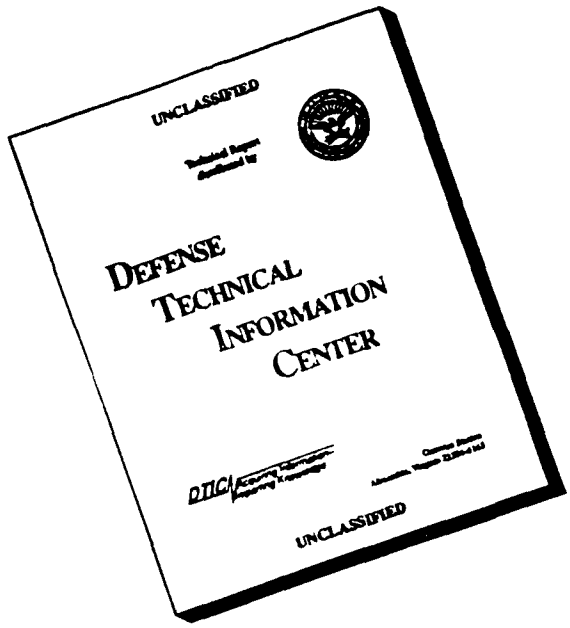
We describe a direct analog implementation of a neural network model of olfactory processing. This model has been shown capable of performing hierarchical clustering as a result of a coactivity-based unsupervised learning rule which is modeled after long-term synaptic potentiation. Network function is statistically based and does not require highly precise weights or other components. We present current-mode circuit designs to implement the required functions in CMOS integrated circuitry, and propose the use of floating-gate MOS transistors for modifiable, nonvolatile interconnections weights. Methods for arrangement of these weights into a sparse pseudorandom interconnection matrix, and for parallel implementation of the learning rule, are described. Test results from functional blocks on first silicon are presented. It is estimated that a network with upwards of 50K weights and with submicrosecond settling times could be built with a conventional CMOS double-poly process and die size.

Published in *Analog Integrated Circuits and Signal Processing*, Vol 2, 1992, pp 35-49.

14. SUBJECT TERMS	15. NUMBER OF PAGES
olfactory synchronous analog granger/lynch	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAME AS REPORT

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

UNCLASSIFIED

21a NAME OF RESPONSIBLE INDIVIDUAL	21b. TELEPHONE *(include Area Code)*	21c. OFFICE SYMBOL
P. A. Shoemaker	(619) 553–5385	Code 552

Analog Integrated Circuits and Signal Processing 2, 297-311 (1992)

A Hierarchical Clustering Network Based on a Model of Olfactory Processing

P.A. SHOEMAKER

Naval Command, Control, and Ocean Surveillance Center, RDT&E Division, San Diego, CA 92152-5000

C.G. HUTCHENS AND S.B. PATIL

Electrical and Computer Engineering Department, Oklahoma State University, Stillwater, OK 74078

Abstract. We describe a direct analog implementation of a neural network model of olfactory processing [44–48]. This model has been shown capable of performing hierarchical clustering as a result of a coactivity-based unsupervised learning rule which is modeled after long-term synaptic potentiation. Network function is statistically based and does not require highly precise weights or other components. We present current-mode circuit designs to implement the required functions in CMOS integrated circuitry, and propose the use of floating-gate MOS transistors for modifiable, nonvolatile interconnection weights. Methods for arrangement of these weights into a sparse pseudo-random interconnection matrix, and for parallel implementation of the learning rule, are described. Test results from functional blocks on first silicon are presented. It is estimated that a network with upwards of 50K weights and with submicrosecond settling times could be built with a conventional CMOS double-poly process and die size.

1. Introduction

In recent years, interest in neural networks and neural-network-like computational models has seen a major resurgence, due at least in part to the prospect of compact and dense implementation of these networks in analog integrated circuit form. A number of widely studied architectures and algorithms are based on adaptations of conventional statistical and numerical techniques which admit parallel network implementations (e.g., multilayer perceptrons with back-propagation learning [1], learning vector quantization [2], and radial basis function or probabilistic neural networks [3, 4]), or on analogy with physical systems (e.g., Hopfield networks [5] and Boltzmann machines [6]). These might be properly termed artificial neural network algorithms, with emphasis on the artificiality, since resemblance to real neural networks (beyond the parallel structure of interconnected processing units) is likely to be either superficial or coincidental. These algorithms have been applied with some success to a number of problems, although studies of them have been conducted almost exclusively in simulations. Much debate has centered on the relative advantages, and even feasibility, of analog versus digital implementations [7, 8]. With the architectures and algorithms that are commonly reported, the precision with which interconnection weights can be represented and the resolution of weight changes during learning are important issues in both the digital and analog cases.

Elucidation of the computational principles used in real nervous systems, on the other hand, has been very limited due to the extreme experimental difficulties encountered in network neuroscience. Understanding of collective function of neural networks in vertebrates is largely limited to sensory structures and early processing, which have been studied in the greatest depth and with the most success; even in these cases, interpretation of the computational principles which are followed is a matter of current research [9–11].

A number of the direct analog implementations of neural networks that have been reported to date consist of building blocks that are suitable for the artificial paradigms; the layered heavily interconnected feedforward architecture epitomized by the multilayer perceptron [12–18] or the reciprocally and symmetrically interconnected architecture described by Hopfield [5] and Cohen and Grossberg [19] are often targeted [20–22]. By way of contrast, some researchers, most notably Mead and co-workers, have attempted to build reasonably faithful analogs of biological neurons or networks [23–29], which are generally early processing structures for sensory input. Mueller and co-workers have reported an intermediate approach with a chipset retaining some notable features of biological neurons but allowing programmable interconnection into general networks [30].

An outstanding problem in analog networks is the practical implementation of learning, which in the neural network field usually comprises some algorithmic procedure for modification of interconnection weights between neuronal analogs in response to stimuli and possibly desired response or other feedback presented to the network. Few implementations reported to date actually include learning of this kind on chip [17, 20, 22]. Implementations of biologically inspired networks are often hardwired [24–26], although a few models with limited adaptive capabilities have been built [27, 28]. A central research issue for implementation of the artificial learning paradigms is the precision with which weight or other parameter changes may be calculated (dependent upon precision of components such as weight circuits) and imposed. A suitable analog medium for long-term storage of weights or other parameters is also a matter of current research; floating-gate MOS or MNOS devices have been proposed for this purpose, and studied by a number of workers [12, 27, 31–36]. The potential due to the charge stored on such a structure could be used to control the conductance of a transistor or transistors in a circuit performing the weighting function. However, the processes by which the stored charge may be altered require either UV irradiation, or high programming voltages to induce Fowler-Nordheim tunneling or hot-carrier injection. In the latter cases particularly, the charging phenomena are very nonlinear and sensitive to geometries and processing parameters [37], and thus it is difficult to conceive of precise modification of analog weights without some kind of local closed-loop control. A few workers have proposed modifications of established algorithms, such as very coarse quantization of weight updates [38, 39], which circumvent the need for imposition of precise weight changes, but the practicability of implementing even these learning rules in parallel in analog circuitry remains to be demonstrated.

In biological neural networks, modulation of synaptic efficacy has long been regarded as a likely mechanism for learning and memory [40], and the phenomenon of long-term potentiation (LTP) as observed in the hippocampus, limbic system, and certain cortical structures is one candidate for this type of mechanism [41–43]. Changes in synaptic strength due to LTP are thought to be rather coarse [43], in contrast with the graded and precise weights and weight changes which are required by the artificial paradigms. How a nervous system might work within such constraints to perform useful computation and to learn effectively is a question whose resolution is stymied by the paucity of information on network-level function within the brain. However, a potentially useful model for olfactory processing has been proposed by Granger, Lynch, and Ambros-Ingerson [44–48] which we believe provides some preliminary answers to questions of this kind. This model deals with the interacting structures of the olfactory bulb (which receives input from the olfactory receptors via the olfactory nerve) and the piriform cortex, as they appear in olfactory mammals such as the rodents and lagomorphs. It was developed to study the function of these structures based on their known anatomy and physiology, and its emergent computational properties, rather than appearing by design, were discovered upon analysis of simulation results. Function is acquired by an unsupervised learning rule, effectively based on coactivity, which models long-term potentiation. Operation is dependent upon the statistical properties of large assemblages of neurons with sparse, combinatorial interconnections and coarse-valued weights.

In this paper, we discuss this model and the features which make it amenable to implementation, and we describe ongoing efforts toward such an implementation in analog CMOS integrated circuitry. The low-resolution weights and coarse, unidirectional weight changes allow a parallel implementation of the learning rule, using floating gates for nonvolatile analog weight storage. Designs of test circuits for macrocells which implement the required functions are presented, and the integration of these macrocells into a complete network is discussed.

2. The Model

The interested reader is referred to the work of Granger et al. for details of the olfactory model [44–48]. The essential features of the model which are relevant to the proposed implementation are summarized as follows. The olfactory bulb receives input from the olfactory receptor neurons in a somewhat topographic fashion: a particular type of receptor cell (i.e., a receptor which responds to particular chemical stimuli) projects its axons along with those of similar cells to a delimited area of the olfactory bulb which is denoted a glomerulus. The aggregate firing rate of these input cells is regarded as the input to the corresponding glomerulus. There are many glomeruli in the olfactory bulb, each associated with a different type of receptor cell, and thus the system input collectively may be regarded as a vector. The input components, which are

excitatory, are first combined with inhibitory feedback signals to be discussed below. The resulting net inputs are subject to nonlinear processing (saturating low and high) as well as a global normalization, mediated by certain inhibitory cells, which limits total bulb activity. The mitral cells, or excitatory neurons within the olfactory bulb, are regarded as two-state or McCulloch-Pitts neurons, which are either quiescent or active. Those within each glomerulus have a range of differing excitation thresholds at which they become active. The normalization contrains the bulb so that only some fraction (on the order of 20% or so) of all mitral cells do in fact become active upon stimulation. The net effect of the processing within the glomeruli is thus as follows: the most significant components of the net input vector are accentuated while many others are suppressed by the constraint on total activity, and the output of each glomerulus is a "thermometer-coded" version of this processed signal, in which the signal intensity is represented by total number of active cells (due to differing thresholds) within the glomerulus.

The outputs of the mitral cells then project to the piriform cortex via the lateral olfactory tract (LOT). Synpases with piriform cells, which are excitatory, are sparse and combinatorial rather than topographic: they appear to be made essentially at random, with a relatively low probability (on the order of 10%). (Piriform cells in the caudal region of the piriform cortex also receive excitatory inputs from cells in the rostral piriform via associational fibers, although this feature will not be discussed in any detail in this paper.) The excitatory piriform cells are arranged in groups or patches, which are defined by strong local inhibition that results in a "winner-take-all" characteristic: only one or a few of the most strongly stimulated cells within each patch reach an active state at any one time. These cells are also modeled as two-state devices. The sparse pattern of winning cells within the patches is regarded as the spatially encoded output of the olfactory bulf/piriform system; these active cells are those which happen to receive a relatively large number of their synapses from active mitral cells. After a burst of activity, piriform cells undergo afterhyperpolarization, which results in a refractory period of negligible or very reduced excitability.

The active piriform cells in turn inhibit the glomeruli of the bulb via another pathway (this is the feedback inhibition which is summed with glomerular inputs). The inhibition is effected by means of synapses which develop according to a correlational or Hebb-type learning rule, resulting in strongest inhibition of those glomeruli most responsible for the firing of "winning" piriform cells.

The reciprocal process of feedforward excitation of the piriform by the olfactory bulb followed by feedback inhibition of the bulb by the piriform is repeated cyclically at the so-called theta rhythm, to which activity in this part of the brain, as well as the animal's sniffing behavior, is synchronized. Feedback inhibition of the bulb during this multiple sampling cumulative. Thus, as the animal sniffs a single odor, the following sequence takes place in the naive network: after the first sniff, the glomeruli with the most significant input components are most strongly inhibited, allowing secondary components to elicit more significant responses from their glomeruli during the next sniff. In subsequent sniffs, these components are also inhibited allowing still weaker components to be expressed, and so on in a hierarchical fashion. At each step in this hierarchy, a novel piriform output code is guaranteed by the refractory state of previously active piriform cells.

Learning in this system, which is modeled after long-term potentiation, is coactivity-based: the weights of excitatory synapses from active mitral cells onto "winning" piriform cells are incremented. Learning is mediated by external inputs from higher cortical regions (i.e., it can be turned on or off). Weights can saturate; when fully potentiated they are larger than naive weights by a factor of only two to three. Learning increments are of constant magnitude and typically represent 5%–10% of the range between naive and fully potentiated weights. LTP, as the name implies, is a long-lasting phenomenon in which measurable weight decay is not observed.

The effect of learning in this model is that the network develops a tendency to cluster its input vectors: the output codes for vectors sufficiently close in the input space become very similar or identical, as the weights associated with piriform cells that have "won" most frequently become larger. Moreoever, the feedback from piriform to bulb then tends to inhibit the glomeruli not simply in proportion to their activity, but rather in relation to the expected activity for the cluster mean. Thus, not only are glomeruli with significant input components suppressed, but in addition, differences between the input vector and the cluster mean tend to be accentuated. The net result is that, during the multisampling process, a hierarchical clustering takes place, in which initial output codes indicate broad class or cluster membership, and subsequent codes, subcluster or narrower class membership. Cluster and subcluster breadth in the input vector space are influenced by the weight increment size, the ratio of saturated to naive weight values, and the data sample

on which the network learns. The essential features of this model have been abstracted and embedded in a somewhat simplified version, whose resemblance to several other unsupervised clustering algorithms has been noted [45, 46].

A number of features of this model are particularly favorable for simple direct implementation. The neuron models are two-state devices, and consequently, four-quadrant multipliers are not required to implement the interconnection weights; in fact, single transistors suffice. However, most crucially, the weights require only low precision, on the order of 3–5 bits, and learning in the network comprises coarse, unidirectional weight changes which take place according to a simple Hebb-type or coactivity-based update rule. Weights saturate as well, and this is a natural feature to be expected of any analog storage medium.

3. Implementation

We propose a direct implementation of this algorithm in the form of a synchronous, analog silicon model in CMOS circuitry. The importance of the theta rhythm for the network function of hierarchical clustering suggests the suitability of an approach which is synchronous or clocked at the highest level of function. External inputs (analogous to inputs from olfactory receptors) would be sampled periodically at an artificial "theta rhythm." For each cycle of this rhythm, there would be two major phases: activation of the bulb and feedforward excitation of the piriform, followed by feedback inhibition of the bulb by the piriform. Between clock cycles, however, computation of neuronal inputs and activitations would be analog, asynchronous, and carried out in parallel. We also propose to implement network learning, with modifiable nonvolatile weights which are updated in parallel according to the Granger/Lynch/Ambros-Ingerson model when network plasticity is desired. Below we discuss the general approach, and then present circuits designed to implement the requisite functions.

3.1. General Approach and Architecture

Following the Granger/Lynch/Ambros-Ingerson model, neuronal analogs in both the bulb and piriform layers are two-state devices. In the bulb, net inputs to the glomeruli are formed by combining positive external input signals with (negative) inhibitory feedback, and these net inputs are then subject nonlinear process and normalization. Within the framework suggested the biological model, we have developed a pair of alt natives for this processing/normalization which are i plementable with closed-loop circuits similar to th used in automatic gain control (AGC). One most clos follows the form given by Ambros-Ingerson [45], c sisting of a vector AGC loop with sigmoidal nonlinc ity acting on each component within the loop, as illu trated in figure 1a. A second includes an AGC lo without the sigmoids, but with a global offset add to each component within the loop such that the larg net input elicits maximal activity from its glomerul This offset is computed by a fast inner loop, as sho in figure 1b. The second scheme may offer some rep sentational advantages, but the relative applicability the two approaches is currently under investigation system-level simulations.

Subsequent to this normalization, the process signals are thermometer-coded by the two-state mitr neuron models in each glomerulus. Individual mitr cell analogs respond with a binary output, indicatir active or inactive.

In the piriform model, subnetworks of neuron analogs are arranged in winner-take-all patches, eac operating with a single global feedback line to achiev patchwide inhibition of "losing" cells. Global feedbac implies that an N-cell patch would be implementabl with complexity of order (N). Such feedback networl have been described by Lazzaro et al. [49].

For "synaptic" weights, we propose the use of analo floating-gate memory in conjunction with a single trans istor weighting element whose conductance is modu lated by charge on the floating gate. Because 10 or fewe distinct synaptic strengths are required for the LO synapses in the Granger/Lynch/Ambros-Ingerson mode [44–48], analog floating gates would seem to pos little risk. Long-term (decades) retention of at least bits of resolution has been estimated by extrapolatio from high-temperature charge-relaxation data on float ing-gate circuits used in an analog neural network im plementation [12].

In the model, the synapses from mitral cells ont piriform cells form a sparse, random interconnection matrix. The approach which we propose to implement this matrix employs a simple one-to-one correspond ence of the number of weighting elements to number of synapses in the model, with mask-programmable connection of input and output lines allowing establish ment of the sparse pseudorandom connectivity. The physical weight matrix is composed of cells containing

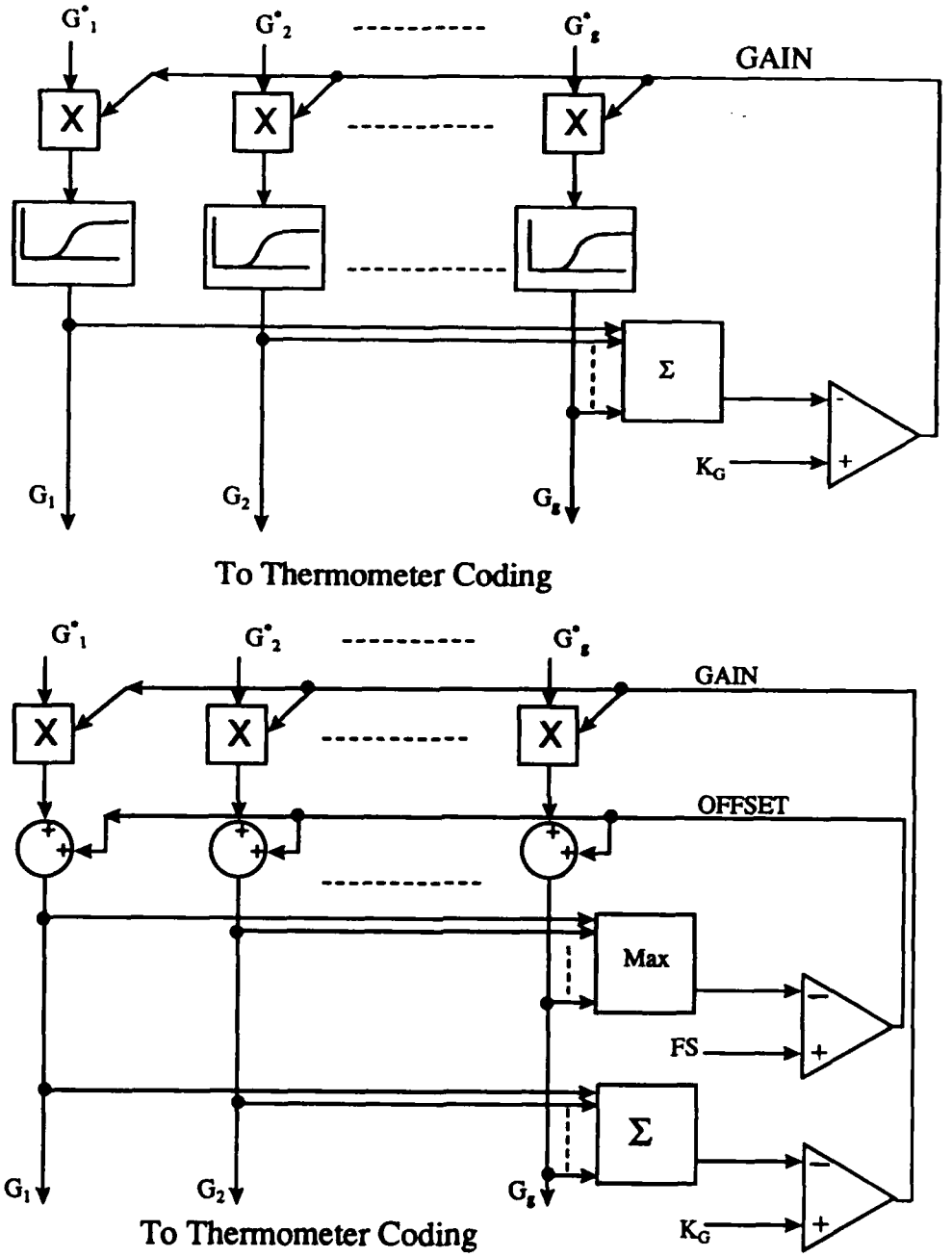


Fig. 1. Schematic diagrams for normalization of net input vectors to hierarchical clustering network. G_i^* represent net input components and G_i normalized input components. (a) Scheme which closely follows the original biological model, with sigmoidal nonlinearity blocks included in the feedback loop. K_G is a reference level corresponding to desired total activation. (b) Scheme which insures that the largest net input component elicits a full-scale response. FS is a reference level corresponding to full-scale activation. Normalized output components are assumed to saturate low at zero.

one or more weighting transistors and the crossing of several mitral output and piriform input lines and interconnections are established at random between pairs of input and output lines within each cell. We consider a prototype for this concept in which a basic weight cell contains two weighting transistors and the crossings of four mitral output lines and five piriform input lines. Any input line may be interconnected with any output line, with the caveat that double interconnection between a given pair of lines is excluded; a connectivity ratio of 1:10 is thus maintained by the use of this cell. The connections are established at layout time by a

macro which generates a randomized list and then places geometries on the appropriate mask layer(s) to establish the interconnections in the layout database. The objective of this approach is to minimize interconnect and routing area and conserve the number of devices required in the interconnection matrix, which factors are of concern [7, 50] in a direct, nonmultiplexed implementation. Assuming scalable design rules, we estimate the area required for this scheme is on the order of one-fifth to one-tenth the area estimate given by Hammerstrom and Means [50] for direct implementation, and which is cited by them as a motivating factor for development of a broadcast multiplexed digital architecture as an alternative to the direct analog approach.

The price paid for the simplicity of the proposed architecture is the forfeiture of a certain degree of statistical independence of the connectivity. For example, three particular LOT lines which pass through the same basic weight cell have zero probability of synapsing onto the same piriform input line, and three piriform lines passing through the cell have zero probability of receiving synaptic input from the same LOT line. Without the constraint imposed by the weight cell, the probability of either of these events is $(1/10)^3$ or 1/1000. However, as a consequence of the central limit theorem, the distribution of active synapses onto the piriform input lines becomes similar to that of the unconstrained interconnection pattern of the original model as the number of LOT lines increases. We have calculated both distributions for LOTs of several hundred lines and mitral activity of 20%, and they are very similar; thus use of the weight cell is not regarded as an important constraint in networks which are sufficiently large, but still of realizable size.

To implement feedback inhibition of the bulb by the piriform, we propose a time-duplex scheme. The original algorithm call for distinct feedback paths from piriform to bulb, with inhibitory synapses trained according to a correlative or Hebb-type learning rule in a developmental phase prior to the application of structured input. However, since these correlations arise in direct consequence of the given connectivity of the LOT synapses, the same effect can be obtained by using the transpose of the LOT weight matrix to compute bulbar inhibition. Physically, this implies that a single weight matrix can be used to compute excitatory bulbar input to piriform, followed by inhibitary currents from piriform feedback to bulb. In the second phase, winning piriform cells would drive the weight matrix, and the output currents would be summed over each glomerulus on the bulb side to obtain the inhibition for that samp or "sniff."

For individual weights, the control logic for t coactivity-based learning rule corresponds to a simp AND function; taken in parallel it may be regarded a Boolean outer product. This can be implemented crossbars running through the weight matrix using si ple switches which are controlled by the neuron sta and which route programming voltages to writing c cuitry for the floating-gate weights.

A block diagram representing an overview of t proposed system is shown in figure 2.

3.2. Circuit Designs

Many of the functions which are required to implem the model as described above may be achieved w well-known analog building blocks. In designing c circuitry, a current-mode approach was adopted reasons of improved bandwidth and noise immuni (Voltage-mode signals are assumed at network inp and outputs, however, for convenience of external int face.) A settling time on the order of several hundr nanoseconds was targeted for feedforward excitatory feedback inhibitory phases of network operatic Current-mode circuits in addition permit a simple so tion to the proposed bidirectional, time-multiplexed u of the weight matrix. Interface is made to the weig matrix on both the mitral and piriform sides via typ two current conveyors (CCII) [51], which act as bidir tional buffer/drivers. In the CCII design shown in figu 3, a folded-cascode differential amplifier is used a gain element for wide bandwidth. Its positive inp serves as the reference (Y) terminal of the convey a class AB output stage (MFN and MFP) coupled the negative input forms the voltage-following (X) t minal, and the current output of this stage is in tu copied to give the current (Z) output of the convey

Two options for the initial processing and normaliz tion of input vector are shown schematically in figu la and lb, as noted in Section 3.1; we describe th salient components below. For multiplication by t global gain in the AGC loop, both simple voltag controlled active loads and a more complex transco ductance multiplier for improved linearity are und consideration. The transconductance multiplier is modified dual-quad circuit. The sigmoid nonlinear of the first preprocessing option is imposed by the c cuit shown in figure 4, in which the basis Θ_G sets t threshold and V_C sets saturation. The input load is practice a complementary series pair of MOSFE

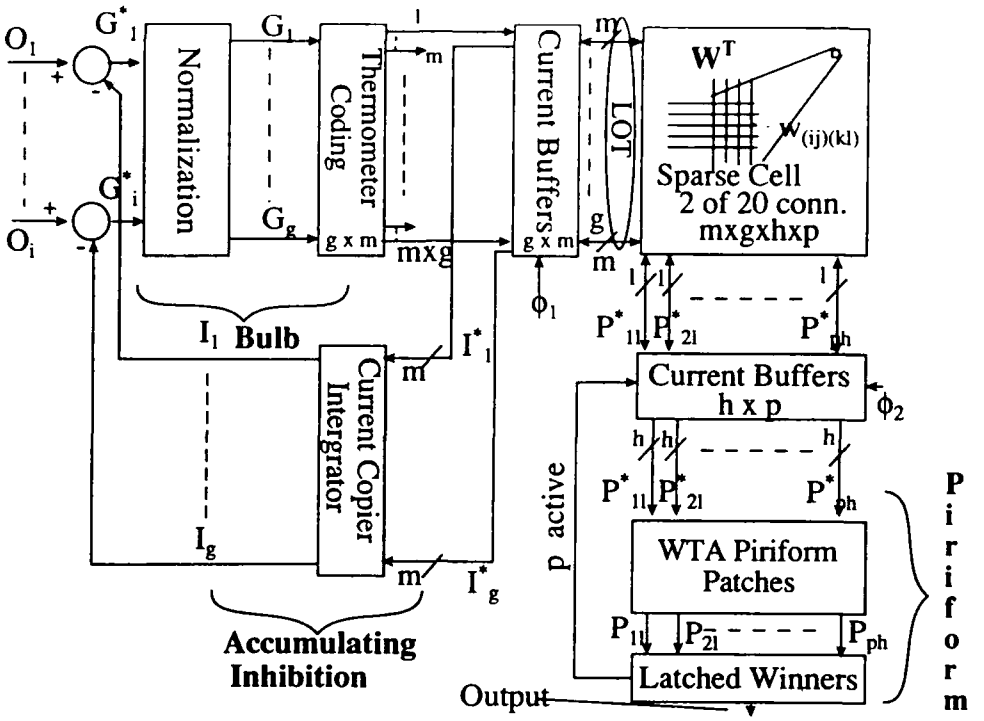


Fig. 2. Overview of proposed system. Integer g indicates number of input components (and bulb glomeruli), m indicates number of levels in the thermometer coding of net inputs, p indicates the number of winner-take-all piriform patches, and h indicates the number of cells per patch. O_i are external inputs, G_i^* are net inputs, G_i are normalized inputs ($i = 1, \ldots, g$), I_j^* are feedback inhibition components ($j = 1, \ldots, m*g$), and I_i are accumulated inhibition for each glomerulus ($i = 1, \ldots, g$). LOT indicates the lateral olfactory tract analog, W^T the transposable weight matrix, and WTA winner-take-all.

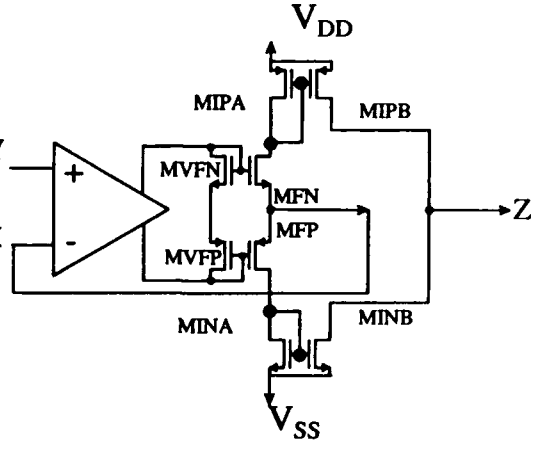


Fig. 3. Schematic of type-two current conveyor (CCII).

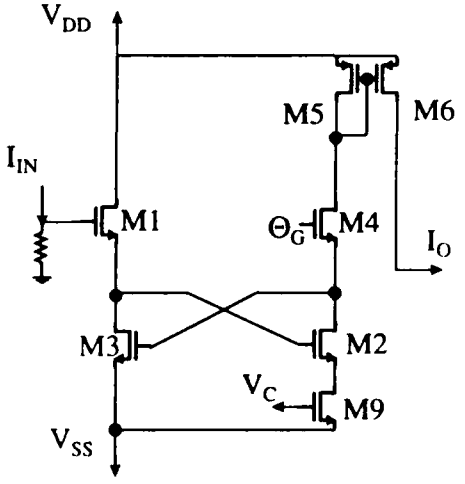


Fig. 4. Sigmoidal nonlinearity circuit with current-mode output. The biases Θ_G and V_C control threshold and saturation characteristics of the function, respectively.

strongly biased in the triode region. The four cross-coupled *n*-channel transistors, *M*1-*M*4, when in saturation, impress the input voltage less the bias Θ_G across nonlinear (saturating) load *M*9, and the current through *M*9 is copied to provide the output of the circuit. In the second option, the offset needed to elicit a full-scale response to largest input component is computed by a fast inner closed-loop circuit as depicted in figure 1b, in which the output of a maximum detection circuit (not depicted) is compared against a full-scale reference. As a gain element in these loops, the folded-cascode differential amplifier embedded in the CCII circuit of figure 3 may be used with the two output terminals connected.

The thermometer-coding function of each glomerulus is achieved with a circuit analogous to the first stage of a parallel analog-to-digital converter, as illustrated in figure 5. A voltage ladder is established by a series of identical capacitors. Full-scale voltage is set globally by equilibrating full-scale input current across a load (again composed of active devices biased strongly in the triode region). In a VLSI network, the full-scale current could be copied and routed to loads in each glomerulus to maintain accuracy. The preprocessed input current for each glomerulus is equilibrated across an identical load and the resulting voltage compared against each step of the voltage ladder by a series of comparators, whose outputs represent the states of the mitral cells within the glomerulus.

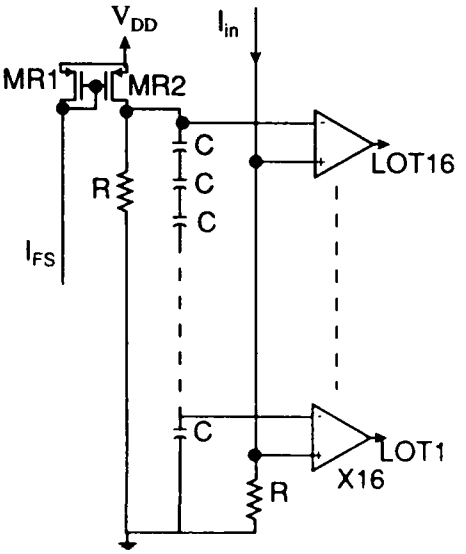


Fig. 5. Thermometer-coding circuit. I_{FS} is a reference current corresponding to full-scale input, and I_{in} is the input current. The two load resistances are composed of active devices in practice, and are identical. LOT1-LOT16 are comparators whose outputs constitute the thermometer encoding of the input.

When the network is in the feedforward mode, th reference (*Y*) input of the current conveyors for acti mitral cells are switched to ground while others a switched to a high reference. On the piriform side, a reference inputs are switched to the high reference. Th *X* terminal voltage follows the *Y* input per normal CC operation.

The weighting elements in the weight matrix eac comprise an individual floating-gate *p*-channel transi tor. The floating gate on the first polysilicon layer capacitively coupled to a "control gate" on the secor polysilicon layer, and the bias applied to the poly-2 co trol gate is used to establish the transconductance corr sponding to the naive weight, when the floating gate uncharged. The bias capacitor is also used to apply programming voltage during learning, to be discusse below. Negative charge on the floating gate increase the transistor transconductance and thus the weight asso ciated with the interconnection. Current flows via th weighting transistors to active mitral cell conveyors fro piriform conveyors, while no appreciable current flow to inactive mitral cell conveyors from piriform cor veyors since both reference inputs are at the same leve

The current (*Z*) outputs of the piriform current cor veyors are routed as inputs to winner-take-all circui which define the piriform patches. The winner-take all circuit depicted in figure 6 operates with global feed back much like the circuit of Lazarro et al. [49], bu is designed for improved sensitivity. It is reset at th beginning of each sniff by transistor *M*5, which distributed in each of the piriform cell analogs, an which discharges the common gate of transistors *M* to V_{ss}. When *M*5 is shut off, this common gate charged by the incoming currents, and when the *M* devices turn on, each begins to sink a portion of th input current for its cell. In all but the cell with th maximum input, the current drawn by *M*1 reaches the exceeds the input current, and the difference curre must be drawn via *M*4. At this transition, the voltag at the input node falls from a threshold above groun to a threshold below. The input node of the single win ner remains near one threshold above ground, with *M* conducting just sufficiently to balance the leakage cur rent from the common gate of the *M*1 transistors. Th voltages at the input nodes are amplified and level shifted by inverters to give the piriform outputs t 0-5 V logic. Transistors *M*2 in figure 6 are cascode included to prevent large swings in the drain voltage of the *M*1 devices.

This analysis assumes that discharge of capacitance at the circuit inputs is fast relative to the charging of th

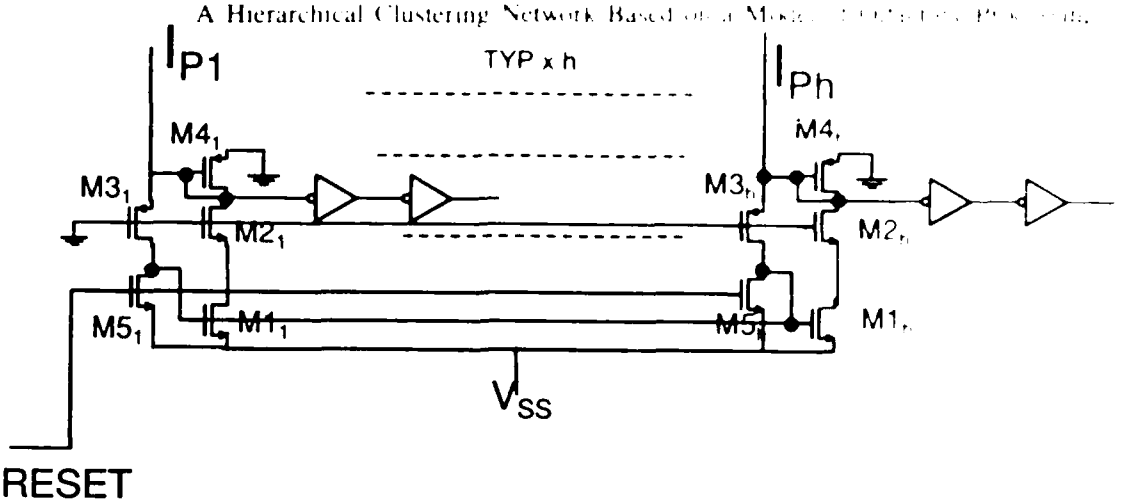


Fig. 6. Winner-take-all circuit for piriform patch with h cells I_{Pi} are input current (i = 1 ... h)

feedback capacitance at the gates of $M1$. If this is not the case, then the gates of $M1$ may overcharge and draw a current greater than the maximum input current during settling, in which case all outputs are pulled low, and remain so while the feedback node is drawn down by leakage off the feedback capacitance, until the $M1$ currents decrease to the maximum input and $M3$ for the cell with maximum input is forced to the edge of conduction.

The sparse pattern of piriform winners from these winner-take-all circuits constitutes the output of the network. Time multiplexing and/or digital encoding would be used in practice to take this data off-chip, in order to limit pin count. To ensure a valid binary code, a digital logic-based tie-resolving circuit has been developed to obtain a single winner from the output of the analog winner-take-all circuit. These circuits are conventional and of secondary concern, and will not be considered further.

After piriform winners are established, the feedback inhibitory phase of the network operation takes place. Piriform states are latched, and the reference inputs for the conveyors of the winning cells are switched to high reference, while those of the losers and of the conveyors on the bulb side are grounded. The output currents of the conveyors for each glomerulus in the bulb are summed and used to determine level of inhibition. Within the general framework of the biological model, several schemes for computation of inhibition are under investigation, ranging from scaling to thresholding of accumulated feedback current before subtraction from external input current. To accumulate feedback over a series of sniffs, a current copier/integrator has been designed as shown schematically in figure 7. The current copier/integrator operates under control of a clock with two (nonoverlapping) phases, the first of which must fall within the feedback phase of the system clock. It is reset before each series of sniffs by discharging hold capacitors CH to V_{dd} or V_{ss}. It includes dynamic current mirrors to enhance the accuracy of the current copying function.

During learning, we propose to exploit simple drain-side hot-electron injection onto the floating gates of the weighting transistors through a gate oxide of usual thickness. This obviates the need for EEPROM or other special processing to implement the floating-gate weights. A scheme for performing coactivity-based updating is outlined as follows. For each mitral output line in the LOT, a corresponding bias line is fabricated which contacts the control gate of every weighting transistor connected to the mitral line. During normal operation, these bias lines are all set at a common bias voltage used to establish the naive weight value. When the weights are to be updated, the bias lines corresponding to active mitral cells are switched to a high-voltage programming line via high-voltage switches, while on the piriform side, the reference inputs of current conveyors for winning piriform cells are strobed to the negative rail, pulling the drains of the weighting transistors for those cells to nearly the same potential. It is assumed that the amplitude of the programming voltage less the lower rail is sufficient to allow injection of some appropriate amount of charge. In this way, the weights interconnecting coactive mitral and piriform cells are incremented. Meanwhile, the reference inputs of the mitral and losing piriform current conveyors are maintained at an intermediate potential such as ground. It is assumed that the programming voltage less the intermediate voltage does not cause injection of significant charge. In addition, the bias lines of inactive mitral cells are held at some potential sufficiently high to maintain the corresponding transistors in a strongly accumulated state and prevent significant channel current in any devices connected to winning piriform cells. In this way, the update rule may be implemented in parallel without drawing large currents.

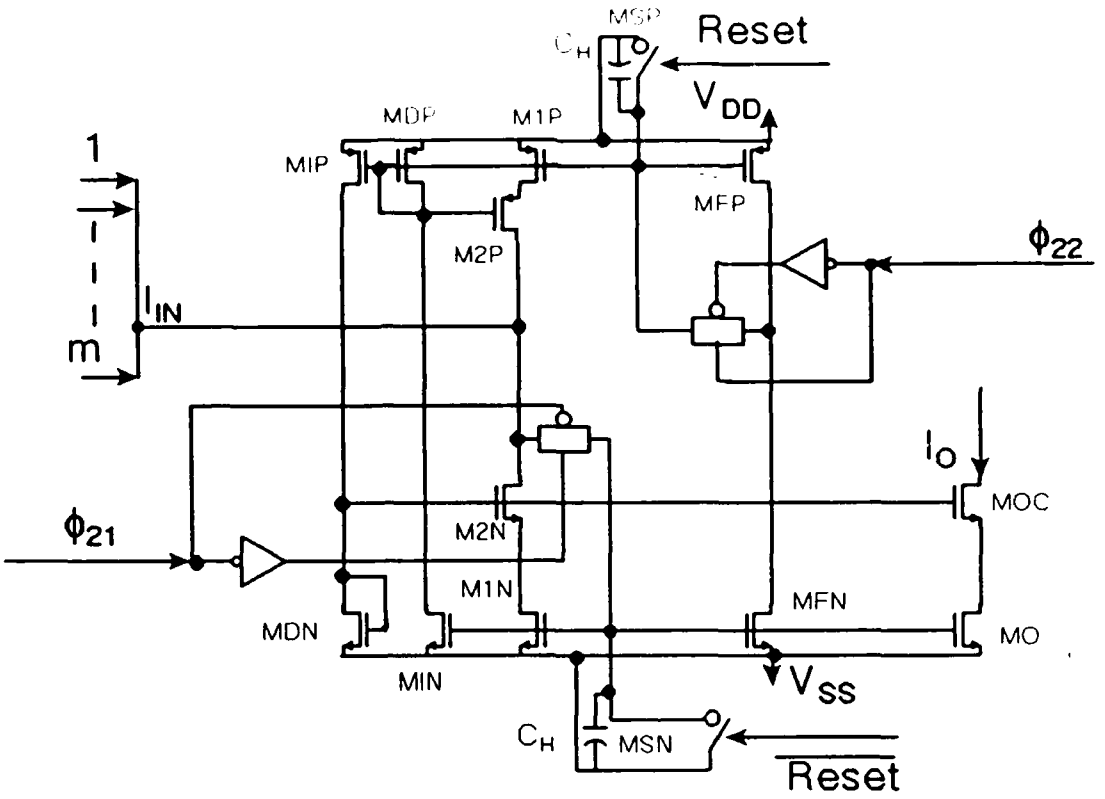


Fig. 7. Current copier/integrator. I_{IN} is input current (summed over m LOT lines connected with a single glomerulus), and I_O is the accumulate output current. ϕ_{21} and ϕ_{22} are nonoverlapping clock phases which control opening and closing of transmission gate switches depicted.

4. Results

Most of the circuits described above were fabricated in a MOSIS 2 μm analog CMOS process, or an Orbit Semiconductor 1.5 μm CMOS process intended primarily for digital applications. Both proceses had double-polysilicon and double-metal layers. Testing performed on these circuits was generally limited to dc functionality as available test resources did not permit full-bandwidth ac or real-time response testing, due primarily to capacitive loading of input and output nodes. Consequently, SPICE simulation results are given to represent the ac frequency or transient response of the circuits. Test results were obtained from either two or three die.

In tests of the CCII circuit, the X output follows the Y reference from -2.5 to 2.5 V under 1-kΩ load, and the Z output stage is capable of tracking the X output current from -2 to 1.75 mA. Simulated unity-gain bandwidth into the 1-kΩ load is in excess of 20 MHz.

Several of the subblocks for the nonlinear normalization circuitry were successfully fabricated and tested. The transconductance multiplier exhibits an rms linearity error (relative to full scale) of 1.7% and a maximum absolute error of 5%. Bandwidth in simulations wit the output loaded by a diode-connected MOSFET ($W/$ = 5 μm/5 μm) is 30 MHz. The sigmoid circuit behave qualitatively as expected in dc tests, with saturation an threshold characteristics controllable by the two bia voltages V_C and Θ_G, respectively. The simulate small-signal bandwidth varies with state but exceed 10 MHz across the range.

The dc transfer characteristic of the thermometer coding circuit is qualitatively as expected, although in put capacitance of the comparators connected to th capacitive ladder contributes to a nonuniformity in ste size of the quantization performed by the circuit. N particular design measures were taken against suc variations as they are believed to be of little significanc as long as quantization is monotonic. In simulations the unloaded comparator outputs respond to a full-scal step input to the thermometer-coding circuit with ris times in the range of 20–150 ns (see figure 8). Respons times are directly related to position of the comparato reference input on the voltage ladder, which determine magnitude of differential drive voltages. The same time order of response occurs in the test circuits wit capacitively loaded outputs.

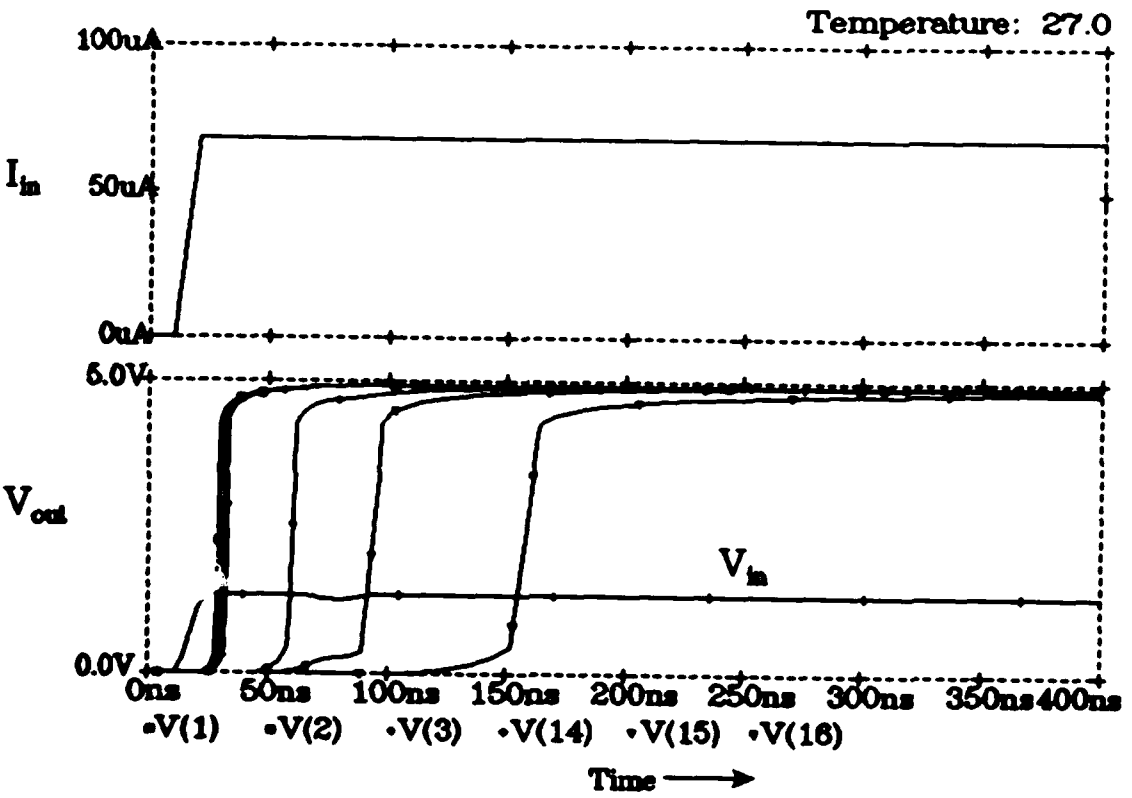


Fig. 8. Simulated transient response of thermometer coding circuit. The trace labeled V_{in} is the input voltage in response to the full-scale step current input in the upper trace. The three leftmost traces on the lower graph are the outputs of comparators at the bottom of the voltage ladder and the rightmost those of comparators at the top of the ladder.

A 32-stage winner-take-all test circuit was fabricated and tested. It was found capable of resolving input currents differing by 1–3 μA at total input levels of 70–140 μA. In eight tests on three circuits, the average resolution was 2.1 μA. As a design target a figure of 5 μA for the current output of a naive weight has been used, so average resolution is to better than half the design current delivered by a single naive weight.

Without added capacitance at the feedback node, the winner-take-all circuit with device geometries as designed has been found to permit overcharging of the feedback node in certain simulated worst-case scenarios. An added capacitance of 2 pF was included in the simulation summarized in figure 9, which depicts time course of response of the circuit after reset in a near-worst-case senario in which the four largest input currents are nearly equal and appreciably larger than the others. The simulation includes no external capacitance at the inputs and outputs. Time to determination of the winner in this case is on the order of 120 ns. Improved performance and elimination of the added capacitance can be achieved by modification of the geometries of the devices in figure 6; in particular, widening of $M1$ will increase both capacitance at the feedback node and the bypass current which discharges capacitance at the input node via $M1$ and $M2$.

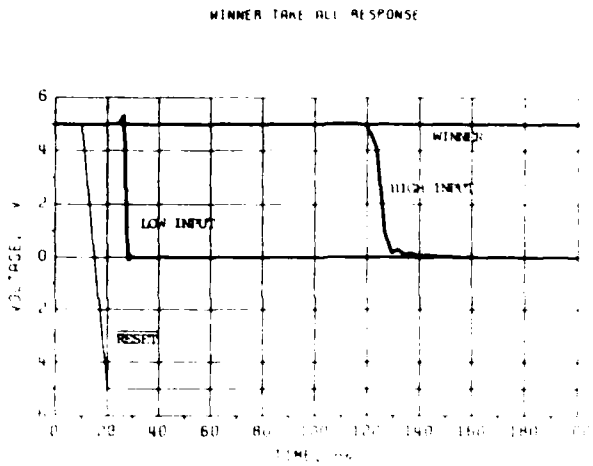


Fig. 9. Simulated transient response of a 32-stage winner-take-all circuit. Twenty-eight inputs were at the low level of 60 μA, three were at the high level of 138 μA, and the winning input was at 140 μA. Examples of three corresponding outputs are shown. Time course of resetting is indicated.

Due to a design/layout error, the current copier/integrator displayed a large copying error (about 70% at 40 μA) at the initial cycle in dc tests. Simulations indicate the circuit to be operable at a clocking speed of 10 MHz.

Floating gate test circuits were fabricated in the 1.5 μm digital process (which had a gate oxide thickness of 25 nm), and tested according to the programming scheme described in Section 3.2. Programming voltages of 17–19.5 V total amplitude (control gate to drain) were used, applied in pulses of several durations and rise times. Positive-going control gate pulses overlapped negative-going drain pulses to prevent channel current flow. Figure 10 depicts shift in transistor threshold voltage (relative to the control gate) observed in one of these tests. These shifts are representative of th potential changes of the floating gate. Useful shifts required microseconds or tens of microseconds of total programming time. Charge relaxation measurem have not been made, although measurable charge does not occur within days at room temperature. Ir dition, in an experiment with 13 V, 1-μs programr pulses applied to the control gate, the drain term was grounded rather than pulsed to −5 V, which null update state in the parallel learning scheme. measurable threshold shift was obtained after 1 ms programming time.

Several unresolved issues remain with regard to of this circuit as a nonvolatile programmable wei One is the strongly nonlinear dependence of charg jection on floating gate potential relative to the dr which decreases as charge builds up on the gate. is reflected in figure 10, in which the abscissa is plc log-scale. The relationship does result in effec saturation of the weight but the uneven increment s during the first few pulses are of concern with re

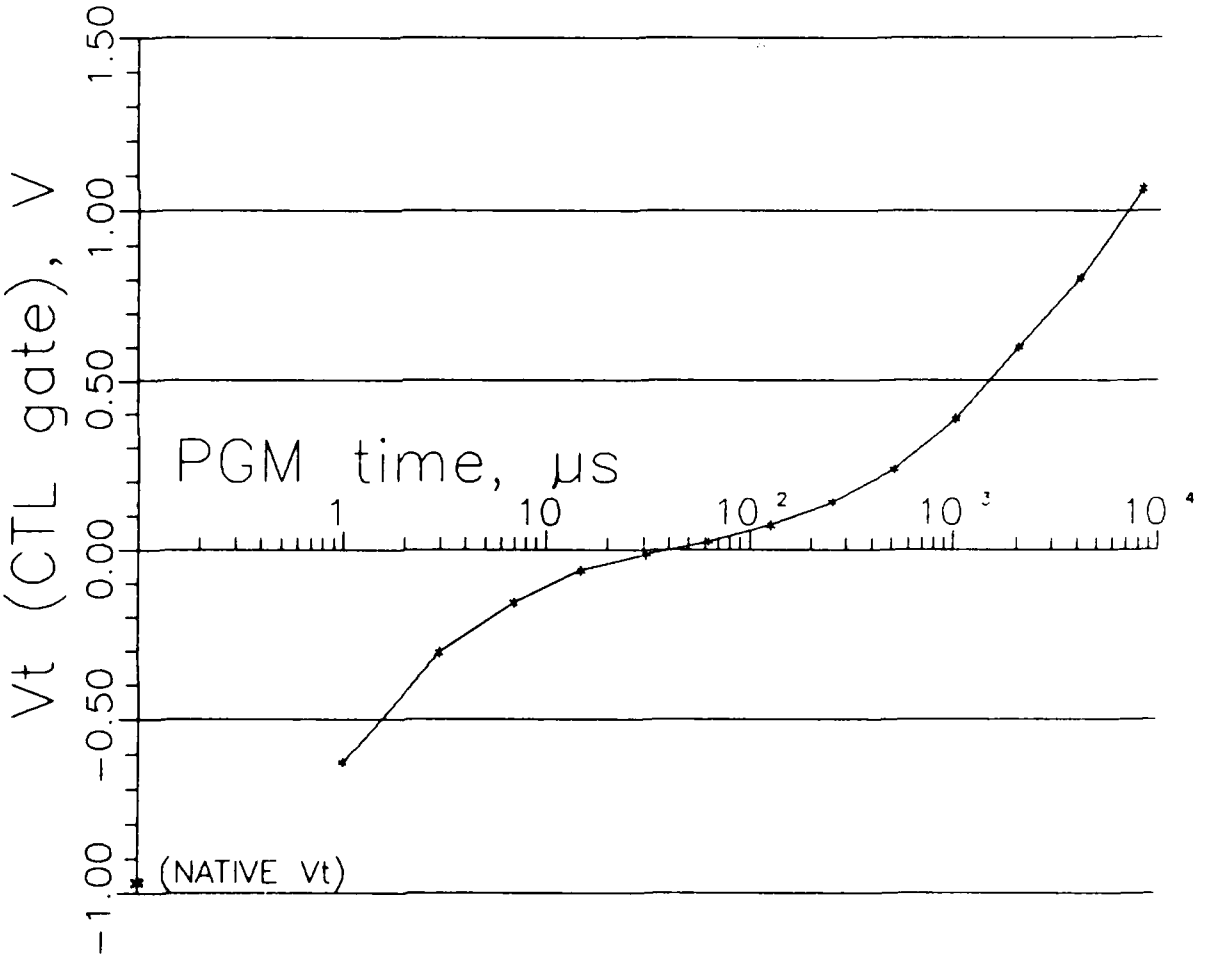


Fig. 10. Threshold voltage shift for a floating gate test circuit subjected to 1-μs, 18-V programming pulses with 120-ns rise times. Thresh is measured relative to the control gate.

to the learning algorithm. Methods of circumventing this problem (e.g., see [52]) are under consideration. In addition, an intermittent rapid (single pulse) charge-up of the floating gate was observed in tests with short (100 ns) overlap of drain pulses by control gate pulses and total programming voltages of 19 V or greater, suggesting a transient junction breakdown or similar phenomenon generating large numbers of hot carriers. The effect was not seen when the overlap was increased to 1 μs, however. Additional experiments are planned in which overlaps, risetimes and pulse widths will be further varied.

5. Conclusions

We have described a model of a neural network which is based upon the known anatomy and physiology of the olfactory bulb and piriform cortex of olfactory mammals [44–48]. This model includes the effects of learning assumed to take place via long-term synaptic potentiation, and it has been shown to be capable of performing hierarchical clustering as a result of this unsupervised learning. Moreover, network function is statistically based and it does not require precise components; in particular, the resolution of the weights needs only be 3–5 bits, and learning is via a simple coactivity-based weight update rule. These characteristics suggest the feasibility of a direct analog implementation; we describe an ongoing effort toward such implementation in CMOS integrated circuitry, which employs current-mode designs, single transistor floating-gate weights, and features parallel on-chip learning. Circuit designs and test results from functional blocks on first silicon are presented. It is estimated that a network with upwards of 50K weights and with submicrosecond settling times could be built with a conventional CMOS double-poly process and die size.

Acknowledgments

This work was supported by the Office of Naval Research. The authors thank Richard Granger and Michael Carlin for their assistance.

References

1. D.E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing, Explorations in the Microstructure of Cognition*, Vol. 1 (D.E. Rumelhart and J.L. McClelland, eds), MIT Press: Cambridge, MA, pp. 318–362, 1986.
2. T. Kohonen, *Self-Organization and Associative Memory*, 2nd ed., Springer-Verlag: Berlin, 1988.
3. D. Specht, "Probabilistic neural networks," *Neural Networks*, Vol. 3, pp. 109–118, 1990.
4. C.L. Scofield and D.L. Reilly, "Into silicon: real time learning in a high density RBF neural network, in *Proc. Int. Joint Conf. Neural Networks*, Seattle, WA, Vol. 1, pp. 551–556, 1991.
5. J.J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. Amer. Acad. Sci.* Vol. 79, pp. 2554–2558, 1982.
6. G.E. Hinton and T.J. Sejnowski, "Learning and relearning in Boltzmann machines," in *Parallel Distributed Processing, Explorations in the Microstructure of Cognition*, Vol. 1 (D.E. Rumelhart and J.L. McClelland, eds.), MIT Press: Cambridge, MA, pp. 282–317, 1986.
7. J. Bailey and D. Hammerstrom, "Why VLSI implmentations of associative VLCNs require connection multiplexing," in *Proc. IEEE Int. Conf. Neural Networks*, San Diego, CA, Vol. 2, pp. 173–180, 1989.
8. S. Morton, "An argument for digital neural nets," Letter to the Editor, *Electronics* May 26, 1988, p. 26.
9. J. Daugman, "Networks for image analysis: motion and texture," in *Proc. Int. Joint Conf. Neural Networks*, Washington, DC, Vol. 1, pp. 189–194, 1989.
10. N. Suga, "Cortical computational maps for auditory imaging," *Neural Networks*, Vol. 3, pp. 3–22, 1990.
11. S.A. Shamma, N. Shen, and P. Gopalaswarmy, "Stereausis: binaural processing without neural delays," *J. Acoust. Soc. Amer.* Vol. 86, pp. 989–1006, 1989.
12. M. Holler, S. Tam, H. Castro, and R. Benson, "An electrically trainable artificial neural network (ETANN) with 10240 'floating gate' synapses," in *Proc. Int. Joint Conf. Neural Networks*, Washington, DC, Vol. 2, pp. 191–196, 1989.
13. D.B. Schwartz, R.E. Howard, and W.E. Hubbard, "A programmable analog neural network chip," *IEEE J. Solid-State Circuits*, Vol. 24, pp. 313–319, 1989.
14. L.D. Jackel, H.P. Graf, and R.E. Howard, "Electronic neural network chips," *Appl. Opt.* Vol. 26, pp. 5077–5080, 1987.
15. F.J. Kub, K.K. Moon, I.A. Mach, and F.M. Long, "Programmable analog vector-matrix multipliers," *IEEE J. Solid-State Circuits* Vol. 25, pp. 207–214, 1990.
16. S.P. Eberhardt, T. Duong, and A.P. Thakoor, "Design of parallel hardware neural network systems from custom analog VLSI 'building block' chips," in *Proc. Int. Joint Conf. Neural Networks*, Washington, DC, Vol. 2, pp. 183–190, 1989.
17. G. Cauwenberghs, C.F. Neugebauer, and A. Yariv, "Analysis and verification of an analog VLSI incremental outer-product learning system," *IEEE Trans. Neural Networks*, Vol. 3, pp. 488–497, 1992.
18. J.B. Lont and W. Guggenbuehl, "Analog CMOS implementation of a multilayer perceptron with nonlinear synapses," *IEEE Trans. Neural Networks* Vol. 3, pp. 457–465, 1992.
19. M.A. Cohen and S. Grossberg, "Absolute stability of global pattern formation and parallel memory storage by competitive neural networks," *IEEE Trans. Syst. Man, Cybernetics*, Vol. SMC-13, pp. 815–826, 1983.
20. J. Alspector, R.B. Allen, V. Hu, and S. Satyanarayana, "Stochastic learning networks and their implementation," in *Proc. IEEE Conf. Neural Infor. Processing Syst.—Natural Synthetic*, Denver, CO, (D.Z. Anderson, ed.), American Institute of Phsycis: New York, pp. 9–21, 1988.

21. D.W. Tank and J.J. Hopfield, "Simple 'neural' optimization networks: and A/D converter, signal decision circuit and a linear programming circuit," *IEEE Trans. Circuits Syst.*, Vol. CAS-33, pp. 534–541, 1986.
22. Y. Arima, K. Mashiko, K. Okada, T. Yamada, A. Maeda, H. Kondoh, and S. Kayano, "A self-learning neural network chip with 125 neurons and 10K self-organization synapses," *IEEE J. Solid-State Circuits*, Vol. 26, pp. 607–611, 1991.
23. C.A. Mead, *Analog VLSI and Neural Systems*, Addison-Wesley: Reading, MA, 1989.
24. C.A. Mead and M.A. Mahowald, "A silicon model of early visual processing," *Neural Networks*, Vol. 1, pp. 91–97, 1988.
25. C.A. Mead, X. Arreguit, and J. Lazzaro, "Analog VLSI model of binaural hearing," *IEEE Trans. Neural Networks*, Vol. 2, pp. 230–236, 1991.
26. A. Moore, J. Allman, and R.M. Goodman, "A real-time neural system for color constancy," *IEEE Trans. Neural Networks*, vol. 2, pp. 234–237, 1991.
27. C.A. Mead, "Neuromorphic electronic systems," *Proc. IEEE*, Vol. 78, pp. 1629–1636, 1990.
28. S.P. DeWeerth and C.A. Mead, "An analog VLSI model of adaptation in the vestibulo-ocular reflex," in *Advances in Neural Information Processing Systems 2* (D. Touretzky, ed.), Morgan-Kaufmann: San Mateo, CA, pp. 742–749, 1990.
29. M. Mahowald and R. Douglas, "A silicon neuron," *Nature*, Vol. 354, pp. 515–518, 1991.
30. P. Mueller, J. van der Spiegel, D. Blackman, T. Chiu, T. Clare, J. Dao, C. Donham, T. Hsieh, and M. Loinaz, "A general purpose analog neurocomputer," in *Proc. Int. Joint Conf. Neural Networks*, Washington, DC, Vol. 2, pp. 177–182, 1989.
31. R.L. Shimabukuro and P.A. Shoemaker, "Circuitry for artificial neural networks with nonvolatile analog memories," in *Proc. IEEE Inter. Symp. Circuits Syst.*, pp. 1217–1220, 1989.
32. V. Hu, A. Kramer, and P.K. Ko, "EEPROMs as analog storage devices for neural nets," First Annual Meeting, INNS, Boston. Abstract appears in *Neural Networks*, Vol. 1, Supp. 1, p. 385, 1988.
33. H.C. Card and W.R. Moore, "Silicon models of associative learning in *Aplysia*," *Neural Networks*, Vol. 3, pp. 333–346, 1990.
34. B.W. Lee, B.J. Sheu, and H. Yang, "Analog floating-gate synapses for general-purpose VLSI neural computation," *IEEE Trans. Circuits Syst.*, Vol. 38, pp. 654–658, 1991.
35. D.A. Durfee and F.S. Shoucair, "Comparison of floating gate neural network memory cells in standard VLSI technology," *IEEE Trans. Neural Networks*, Vol. 3, pp. 347–353, 1992.
36. J.L. Meador, A. Wu, C. Cole, N. Nintunze, and P. Chintrakulchai, "Programmable impulse neural circuits," *IEEE Trans. Neural Networks*, Vol. 2, pp. 101–109, 1991.
37. S.M. Sze, *Physics of Semiconductor Devices*, Wiley: New York, 1981.
38. P.A. Shoemaker, M.J. Carlin, and R.L. Shimabukuro, "Back-propagation learning with trinary quantization of weight updates," *Neural Networks*, Vol. 4, pp. 231–241, 1991.
39. C. Peterson and E. Hartman, "Explorations of the mean field theory learning algorithm," *Neural Networks*, Vol. 2, pp. 475–494, 1989.
40. D. O. Hebb, *The Organization of Behavior*, Wiley: New York, 1949.
41. T.V.P. Bliss and A. Gardner-Medwin, "A long-lasting potentiation of synaptic transmission in the dentate area of the unanesthetized rabbit following stimulation of the perforant path," *J. Physiology London*, Vol. 232, pp. 357–374, 1983.
42. R.J. Racine, N.W. Milgram, and S. Hafner, "Long-term tiation phenomena in the rat limbic forebrain," *Brain Res* Vol. 260, pp. 217–231, 1983.
43. G. Lynch, *Synapses, Circuits, and the Beginnings of M.* MIT Press: Cambridge, MA, 1986.
44. J. Ambros-Ingerson, R. Granger, and G. Lynch, "Sim of paleocortex performs hierarchical clustering," *Scienc* 247, pp. 1344–1348, 1990.
45. J. Ambros-Ingerson, "Computational properties and beh expression of cortical-peripheral interactions suggested by a of the olfactory bulb and cortex," Ph.D. Dissertation, U sity of California, Irvine, 1990.
46. R. Granger, J.A. Ambros-Ingerson, P. Anton, and G. "Unsupervised perceptual learning: a paleocortical mod *Connectionist Modeling and Brain Function*, Chap. 5 (S son and C. Olsen, eds.), MIT Press: Cambridge, MA,
47. R. Granger, J.A. Ambros-Ingerson, and G. Lynch, "Der of encoding characteristics of layer II cerebral cort *Cognitive Neurosci.*, Vol. 1, pp. 61–87, 1989.
48. G. Lynch and R. Granger, "Simulation and analysis of a cortical network," *Psychol. Learning Motivation*, Vol. 2 205–241, 1989.
49. J. Lazzaro, S. Rychebusch, M.A. Mahowald, and C.A. "Winner-take-all networks of $O(N)$ complexity," Californi tute of Technology, Technical Report Caltech-CS-TR-21-88
50. D. Hammerstrom and E. Means, "System design for a generation neurocomputer," in *Proc. Int. Conf. Neural Ne* Washington, DC, Vol. 2, pp. 80–83, 1990.
51. A. S. Sedra and G.W. Roberts, "Current conveyor theo practice," in *Analogue IC Design: The Current-Mode App* Chap. 3 (C. Toumazou, F.J. Lidgey, and D.G. Haigh Peregrinus: London, 1990.
52. O. Fujita, Y. Amemiya, and A. Iwata, "Characteristics of f gate device as an analogue memory for neural networks," *tron. Lett.*, Vol. 27, pp. 924–926, 1991.

Patrick Shoemaker was born in California in November, 19 completed undergraduate and graduate curricula in bioengi at the University of California at San Diego, from which he re the Ph.D. degree in 1984. His areas of study included biomec and neuroscience. Since 1984, he has been with the U.S. Nava mand, Control, and Ocean Surveillance Center, where his wo rently involves analog integrated circuit implementation and a tions of neural network models. He has published papers in network, electronics, and biological/bioengineering journa holds three patents with three pending. Dr. Shoemaker is a n of the IEEE and the International Neural Network Society

Chris Hutchens received the B.S. and M.S. degrees in electrical engineering from South Dakota State University, Brookings, SD, and the Ph.D. degree from the University of Missouri-Columbia in 1979. He is currently on the faculty of Oklahoma State University, Stillwater, OK, where his interests are analog CMOS VLSI circuits, piezoelectric transducers, and bioengineering. Dr. Hutchens has worked at Naval Command, Control, and Ocean Surveillance Center for the past three summers on electronic neural networks, consulted for AMOCO, Tulsa, OK, and is a certified clinical engineer. He is a member of several IEEE societies and Eta Kappa Nu.

Sanjay B. Patil was born in India. He received the D.E.E. and B.S. degrees in electrical engineering from the Government Polytechnic, Yeotmal, and College of Engineering, Poona, both in India, in 1984 and 1987 respectively. Currently he is a graduate student at Oklahoma State University. His reserach involves analog VLSI models for neural networks. His research articles have appeared in the Oklahoma Symposium on Artificial Intelligence.